SIFpackets User Guide

Freddy Cliquet
LINA, UMR CNRS 6241 Universit de Nantes
2 rue de la Houssinire, 44322 Nantes, Cedex 03, France
freddy.cliquet@univ-nantes.fr

Guillaume Fertin
LINA, UMR CNRS 6241 Universit de Nantes
2 rue de la Houssinire, 44322 Nantes, Cedex 03, France
guillaume.fertin@univ-nantes.fr

Irena Rusu
LINA, UMR CNRS 6241 Universit de Nantes
2 rue de la Houssinire, 44322 Nantes, Cedex 03, France
irena.rusu@univ-nantes.fr

Dominique Tessier
UR1268 BIA, INRA
Rue de la Graudire, BP 71627, 44316 Nantes, France
dominique.tessier@nantes.inra.fr

April 28, 2010

Contents

1 Introduction

The SIFpackets application can perform comparisons between experimental tandem mass spectra (MS/MS spectra) and proteins databank. This comparison is done using the algorithm PacketSpectralAlignment (PSA) developed by Cliquet et al. in [1] and the framework detailed in [2]. `PSA` allows to identify in spite of modifications (insertion, suppression or substitution of one or several amino acids). This is especially useful to search homologous proteins in related organisms.

2 Installation

The installation folder of PacketSpectralAlignment (`PSA`) contain several directories to store all the data needed and the output files.

- `data` contains all the data used by the application
  - `databank` contains the databank to which the spectra will be compared (both fasta and prepared for `PSA` databanks)
    - `fasta` contains the fasta of the databanks
    - `matrices` contains the PAM matrix to allow the creation of modifications inside databanks
  - `matrices` contains precomputed alignment used for the AUC computation
  - `output` contains all the files coming from the comparisons of the spectra
    - `R` contains special files created in order to compute ROC curve using the R language
  - `spectrum` contains the experimental spectra

3 Functionality

3.1 Databank preparation

The Databank preparation functionality is necessary in order to compare experimental `MS/MS` Spectra to a protein databank. This is here that the databank will be virtually digested by an enzyme. Currently, only the trypsin

can be used. The FASTA file of the databank to use is needed and must be place directly in the appropriate folder :
`<Application Location>/data/databank/fasta`.
This tool can be access in the menubar (`Tools` → `Databank preparation`).

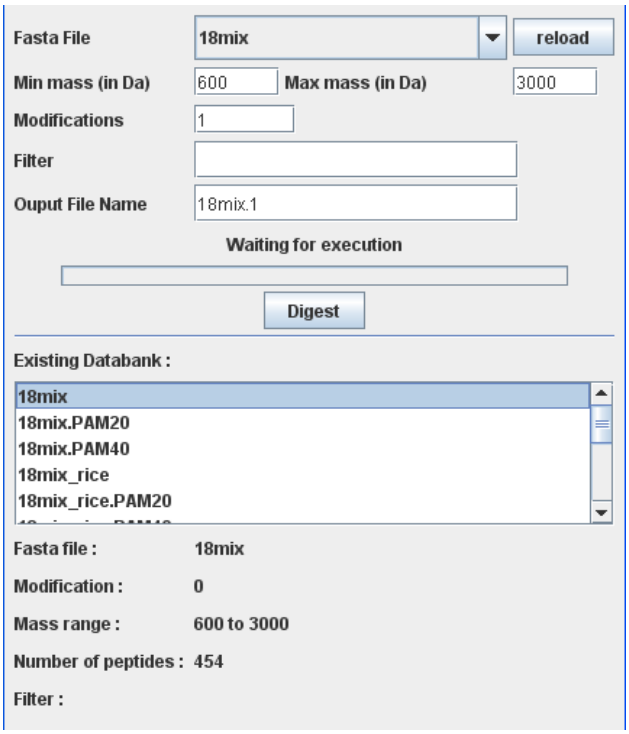

Figure 1: Databank Preparation window.

In the *Databank Preparation* Window (Figure 1), there are several options available. They are listed and explained here:

- The *Fasta File* list allows to chose the databank FASTA file (files appears in the list without the extension `.fasta`). It is possible to reload the content of the list in the case that a new file have been copied in the `fasta` directory by clicking on the reload button.

- The *Min mass* et *Max mass* fields allow to select the range within we will conserve the peptides after the digestion. This mass usually depends on the tolerance of the mass spectrometer used. The default

value (min = 600 Da and max = 3000 Da) are common values for `Q-TOF` mass spectrometers. A wrong estimation of these two parameters can cause some problems, if they are too restrictive (high *Min mass* and low *Max mass*) the application may miss expected peptides, and on the contrary, giving a low *Min mass* value or a high *Max mass* will allow too much comparison (and no good results are expected from these additional peptides).

- The field *Modifications* allows to modify the peptides during the digestion. This possibility is for a purpose of tests, to verify the ability of the method to identify peptides in spite of modifications. The method can either induce a fixed amount of modification per peptide, using an equal probability to each amino acids to mutate into another randomly chosen amino acid, either induce mutation using PAM matrices as defined by Dayhoff. The value of this fields can be either an integer (being the exact number of modification added in each peptide), or a string representing the PAM matrix to use (e.g. "`PAM10`" is a correct value). The default value ("0") means that no modification are applied to the peptides.

- The field *Filter* can be used to filter proteins using their names. Only those with the given string will be kept. It could be useful in some cases, in order to keep only one organism, or even one chromosome from a databank. An empty field means that no filter is applied.

- The field *Output File* Name allow to enter the name used to store the prepared databank. A name is suggested by default, it consists of the FASTA file name concatenated with the modifications applied.

- The *Digest* button start the preparation process with the chosen parameters. A progression bar allows to follow the process.

- The *lower part of the window* allow to look at the databank that were previously prepared. Selecting one in the *Existing Databank* list allow to look at the parameters used to create it, and at the number of peptides composing it.

3.2 Comparison of Spectra against a Databank

The comparison method PacketSpectralAlignment is available through the menubar (`Tools` → `Search Spectra`). This is the tool that will compare all the spectra from an experimental dataset with all the candidates peptides from the chosen databank. Several options are available here:

- The *Spectra File* can be chosen in a list. All recognized spectra files must be stored in the following directory:
  `<Application Location>/data/databank/fasta`.
  The application supports now two different input format, the first one is the Mascot Generic Format (`MGF`) and the second one is the `XML` format from *ProteinLynx*.

- The *Databank File* list allows to select the databank to search on. These databanks have been created using the *Databank preparation tool* (subsection 3.1).

- The field *Output file name* contains the filename used to store the results files. A default name composed of a concatenation of the name with the databank file name is proposed.

- The field *Number of Core* allows to set the number of core to use on the computer, thus allowing the creation of the same number of thread. The default value correspond to the total number of threads of the current computer. This value must be lower if it is needed to run other application while searching spectra or if the computer has memory problems.

- The *K parameter* corresponds to the number of modifications tolerated per 100 Daltons. The default value (0.15) corresponds to an average of one modification per 6 amino acids.

- The *Delta parameter* represents the maximal percentage of tolerated mass difference between the modified and unmodified peptide. The default value, 20, means that if there is more than 20% mass difference between an experimental peptide (given by the parent mass of the experimental spectrum) and a databank peptide, the comparison is not done, and the similarity score is assumed to be 0. The higher the Delta value is, the slower the method will be. This value is a compromise

between speed and quality, and highly depend on the average number of modifications expected.

Our method apply several preprocessing method on the experimental spectra. At first a filtering on the peaks to eliminate noise, by sliding a window on the spectra and keeping only the most intense peaks for each positions of this window. Then, the method compute what we have called possible positions (PP). These are position inside a spectrum that will probably lead to a high scoring alignment with a *packet*. These position are filter at first with a threshold (the minimal score given by a position, and then by a sliding window to keep only the most promising in a certain area of the spectra. It is possible to change all of these parameters:

- The *filter window width* corresponds to the width of the window used to filter the peaks. The default value, 110, matches the average size of an amino acid.
- The *filter window quantity* parameter is the number of peaks kept for each position of the window. The default value, 6, means that we expect to have on average 6 peaks per amino acids. A lower value will reduce the number of peaks other than b and y inside the spectrum reducing the effectiveness of `PSA`, increasing it will lead to keep lot of noise and thus to a much slower method.
- The threshold PP parameter is the minimal score a packet must obtain to consider an alignment position. The default value, 8.00, means that at least a b-ion or a y-ion must be used in the alignment to consider it. A high value will lead to miss lots of positions and thus to a bad scoring alignment, and a low value will only keep lot of wrong PP.
- The *window width to filter PP* parameter is the width of the window used to filter the PP. The default value, 110, matches the average size of an amino acid.
- The *window quantity to filter PP* parameter is the number of PP to keeps for each position of the sliding window. By default, 6 PP are kept for each position. To many positions will strongly slower the method, and to few will speed it up but will also decrease quality. This

parameter strongly relies on the *quality* of the experimental spectra, and on having a *model of packet* perfectly adapted to the data.

The window has a *batch list* at the bottom (left side). When a task is correctly parameter, it must be added to the *Configuration to run* list with the *add* button. Several task may be added. The remove button will delete the selected task from this list. The launch comparison button will starting processing the tasks from this list (top to bottom) showing the progression with the progression bar.

On the bottom right side is the Existing results list to look at already existing files.

3.3 Analyzing the results

The main window of the application allows to analyze results of the search comparison. It is possible to load the existing result file by selecting it in the *Results to visualize* list (if the element does not appear in the list, it is possible to *reload* the list) and clicking on the *Visualize Results* button. Loading results may take some time since the method will try to identify proteins based on the already identified peptides. On the upper part of the screen will be display some data concerning the analysis (such as the spectrum file and databank file used, or the executions parameters).

Then, this tool comports several possibilities: identification of proteins, computing Area under the ROC Curve, and Visualizing Spectra.

3.3.1 From peptides to Proteins

All the proteins identified by at least one identified peptide from the search are listed in the *Proteins* list. Each protein has an associated score (the score is between 0 and 100), the higher this score is, the more chances there is that the protein is correctly identified. Selecting a protein gives its full name (*Protein name* field) and its score with some details: *score complete* that depends on the number of pf spectra used to identify the protein, and *score distinct* that depends on the number of different peptides identified in the protein.

In additon, when a protein is selected, the *Peptides* list shows all the couples peptide/spectrum used to identify the protein. Each couple, when selected, shows in the lower part of the window some informations:

- The *Peptide id* is the id of the peptide in the databank file.
- The *Spectrum id* is the id of the spectrum in the result file.
- The *PSA score* is the score given by the comparison method `PSA`.
- The *peptide sequence* is the peptide sequence that matches the best with the spectrum.
- The *Is low complexity* boolean field gives a complexity information on the peptide sequence. If the sequence contains a string repetition of Glycine, it is consider as low complexity and thus will have lower the peptide score.
- The *peptide score* is the contribution to the score of the protein for this couple spectrum/peptide. It is computed using several factor such as the length of the peptide sequence, and its complexity.
- The *Mass difference* field gives the difference of mass between the experimental spectrum and the theoretical one. If there are no modification, this difference should be close to 0 (it is possible to see a very small variation due to the mass spectrometer precision). If there are modifications, this mass difference can give informations on the modification (much more useful information are available, see the Visualizing spectra part 3.3.3).

3.3.2 Area Under the ROC Curve

In some very particular cases, the visualization tool can compute the Area Under the ROC Curve (AUC) a value that attest the quality of the results. This is only possible if the good association peptide/spectrum is known for for each spectrum of the dataset. It is possible to do so if a **peptide prophet** result file exists and is place with the spectra file. Right now it only works for spectra in the `MGF` file format.

3.3.3 Visualizing spectra

When an element from the peptide list is selected, it is possible to visualize the experimental spectrum and the theoretical one together to see how they

match. This functionality is particularly interesting in presence of modifications, especially as it has been designed to allow the live modification of the peptide to adjust it on the experimental spectrum.

The visualization window contains sevral elements of interest :

- The first one is the list of modifications that allow the better match according to PSA. Each element of this list correspond to one modification and contains at first the localization of the modification (in Daltons, not the number of the amino acid), then the new mass of the element found in this place and then its previous mass. Using those two masses, it is possible to know (at least to have an idea) of what was the element, and more importantly, of what it could be after the modification.
- Then a field with the matching unmodified theoretical peptide. The peptide from this field can be modified, and the spectrum updated using the Refresh Spectrum button. This field supports a particular format :
  - All peptides coded using the one-letter code of each amino acids is supported.
  - An unknown element of known mass can be created using the notation $\_x\_$ where x is its mass in Daltons. This represents an element, so a fragmentation is expected immediately before and after it.
  - The mass of an amino acid can be changed by adding $+x$ or $-x$ just after it, to add or subtract x Daltons. For example, in AM+16GV, the "+16" will increase the mass of the methionine (M) of 16 Daltons (corresponding to an oxidation).
- Several spectra representations are then proposed. The blue peaks represent N-terminal ions, and the red ones represent C-terminal ions.
  - The first spectrum represents the theoretical spectrum from the peptide entered in the field just above.
  - The second one represents the experimental spectrum (after application of symmetry, thus there is a duplication of the peaks).

  - The third one represents the experimental spectrum, with in green the peaks that are aligned with the theoretical spectra according to PSA. (Be careful, symmetry has been applied on this spectrum, thus all blue peaks have a red duplicate one. in most of the cases, only one of them will be aligned. Considering this, not having all the high intensity peaks aligned could be normal.)
  - The last representation is a plot of the possible positions used by PSA. The blue ones represents positions where packets are effectively aligned. The higher a possible position is in this graph, the better the score of the alignment will be.

- Finally, a reminder of the alignment score is shown at the bottom of the screen, this score is updated with the change of the peptide in the field.

4 Parameters

Default parameters are frequently suggested, they have been fixed using several experimentation on real sets of data coming from a Quadrupole Time-of-flight (`Q-TOF`) mass spectrometer. They must give good results, but may eventually depend on the mass spectrometer you use. In addition, right now the only model of packet used to align spectra relies on `Q-TOF` data.

References

[1] F. Cliquet, G. Fertin, I. Rusu, and D. Tessier. Comparison of spectra in unsequenced species. In *Proc. 4th Brazilian Symposium on Bioinformatics (BSB 2009)*, volume 5676, pages 24–35, 2009.

[2] F. Cliquet, G. Fertin, I. Rusu, and D. Tessier. Proper alignment of ms/ms spectra from unsequenced species. In *BIOCOMP*, 2010.